Attachment 9: Identifying Sample Size for Water Filter Performance Assessment (Approved)

1.0 Sample Size Water Filter Testing

The performance of water filters will be tested by comparing unfiltered and filtered water samples at the tap. A key performance metric is the proportion of filters expected to reduce Pb concentrations from levels exceeding 15 ppb to less than the 15 ppb screening level. It is desired to identify a number of households to sample so that the MDHHS can provide a high degree of confidence that 95% of filters are effective. To develop such a test, a number of households with elevated water Pb concentrations must be identified and tested. The number of households that should be tested is a function of the prevalence of elevated Pb concentrations in the population under study, as well as the degree to which water concentrations are elevated. Manufacturers technical specifications suggests that the filters should reduce Pb concentrations by 99.7%, so from a manufacturers perspective a successful test would be one that resulted in at least a 99.7% reduction in filtered water and Pb lead levels in filtered water that does not exceed 5 ppb. However, if unfiltered concentrations exceed 150 ppb, then filtered water would not be safe for drinking with 99.7% reduction per manufacturers tests. Nonetheless for this study, a properly functioning filter will be those filters for which either filtered water has at least 99.7% lower concentrations, or for taps with unfiltered water less than 150 ppb, filtered water is reduced to less than 5 ppb. In cases where Pb concentrations exceed 150 ppb, alternative procedures, such as flushing and then filtering may be needed to bring concentrations to less than 5 ppb. The following results summarize the number of samples necessary to achieve a lower confidence limit on the percentage effective to be at least 99.7% with 95% confidence. The results are developed for a range of sample sizes from 20 to 500 and assuming a range of failures of 0 to 4 (i.e. not reduction by 99.7% or not reducing to less than 5 ppb). The greater the number of failures, the greater the number of total samples needed to achieve the lower confidence limit exceeding 95%. It is anticipated that MDHHS would enter into the study assuming a low sample failure rate of 0 or 1 potentially. Greater numbers of failures in a small sample size would most likely lead to the conclusion that the filters lack effectiveness, as opposed to considering a larger experiment.

1.1 Methods

The generalized linear model provides a robust approach for estimating means and testing hypotheses for right skewed data. The model is based on the assumption of a gamma distribution for data with variance proportional to mean. The model formulation is a log-linear model where the logarithm of expected value of group means is assumed to be linearly related to parameters of interest including discrete variables such as service line type, or continuous variables such as hardness or PH. The model formulation is

$$\log(mean\ Pb_i) = \beta_0 + \beta_1 \times X_1 + \beta_2 \times X_2 + \cdots + \beta_p \times X_p.$$

Where the variables (X_i) represent processes of interest. For example if each of the $X_i$are defined to be 0 for lead service lines and 1.0 for the other service line types, then the regression coefficients β_i represent the log-ratio of Pb concentration in the i[th] type relative to Pb concentrations in water from lead service lines.

1.1.1 Minimum Detectable Ratios

Because of the log-linear relationship, tests of significance of the regression coefficients are equivalent to tests of the null hypothesis that the ratio of concentrations between groups is 1.0.

$$H_0\colon \beta_i = 0 \text{ is equivalent to testing } H_0\colon \frac{Pb_i}{Pb_0} = e^{\beta_i} = 1.0$$

So for determining sample size it is only necessary to identify a number of samples needed to estimate $\beta_i$with reasonable precision. The test of hypothesis is given by comparing a T statistic to a two sided 95% critical value which for moderate to large sample sizes is approximately 1.96.

$$T = \frac{\beta_i}{SE(\beta_i)}$$

We reject the null hypothesis when $|T| > 1.96$ or equivalently when $R = \frac{Pb_i}{Pb_0} > e^{1.96\times SE(\beta_i)}$. Defined in this way, R is the smallest ratio of Pb concentration that would be statistically differentiable from 1.0, as a function of the standard error. To arrive at a reasonable sample size this smallest-detectable ratio was plotted against sample size, so that for planning purposes one can understand the ratio of concentrations that will be precisely estimated in the study design.

1.1.2 Precision of the Mean Estimate

A confidence interval for the mean of any of the groups is found by calculating an interval for model parameters (which are in log-scale) followed by a transformation to concentration scale. For the mean of the ith group an approximate 95% confidence interval in log-scale is

$$\beta_0 + \beta_i \pm 1.96 \times SE(\beta_0 + \beta_i)$$

and the upper and lower limits in concentration scale are obtained by exponentiating the log-scale limits.

$$UCL95 = e^{\beta_0+\beta_i+1.96\times SE(\beta_0+\beta_i)}$$

$$LCL95 = e^{\beta_0+\beta_i-1.96\times SE(\beta_0+\beta_i)}$$

Because standard errors are calculated on the natural log scale, the standard error of the mean is multiplicative in concentration scale, so a simple way to express precision of the mean is simply to subtract the estimated mean from the upper confidence limit, or analogously to subtract the lower confidence limit from the mean. The upper and lower precision are not symmetric in concentration scale, but are symmetric in log-scale. In what follows, we simulated upper interval width for mean concentrations of 1, 5, 15, 30 and 100, for sample sizes of 5, 10, 15, 20, 25, 30, 40, and 50 per group.

1.1.3 Simulating Precision and Power

To arrive at this relationship, we estimated the coefficients of a generalized linear model based on the pilot Pb data from Highland Park, and then simulated many synthetic data sets with similar statistical properties, but with varying sample sizes ranging from 5 to 50 samples per group. While we estimated group means based on service line type, the sample size requirements developed here are general to any grouping variables with similar or lesser within group variability relative to the service line types. In what follows, we simulated upper interval width for mean concentrations of 1, 5, 15, 30 and 100, for sample sizes of 5, 10, 15, 20, 25, 30, 40, and 50 per group.

1.2 Sample Size Water Filter Testing

The number of samples necessary to achieve a lower confidence limit greater than 95% increases with the number of failures. If no failures are observed, a minimum of 80 samples would be adequate to conclude that the effectiveness rate is 95% or greater with 95% level of confidence ([REF _Ref20432787 \h * MERGEFORMAT][REF _Ref23938212 \h * MERGEFORMAT]). Conversely, with 4 failures at least 200 comparisons of filtered and unfiltered water would be required to establish 95% effectiveness with 95% confidence. It is also important to note that for fewer than 60 samples it would not be possible to establish 95% effectiveness with 95% level of confidence.

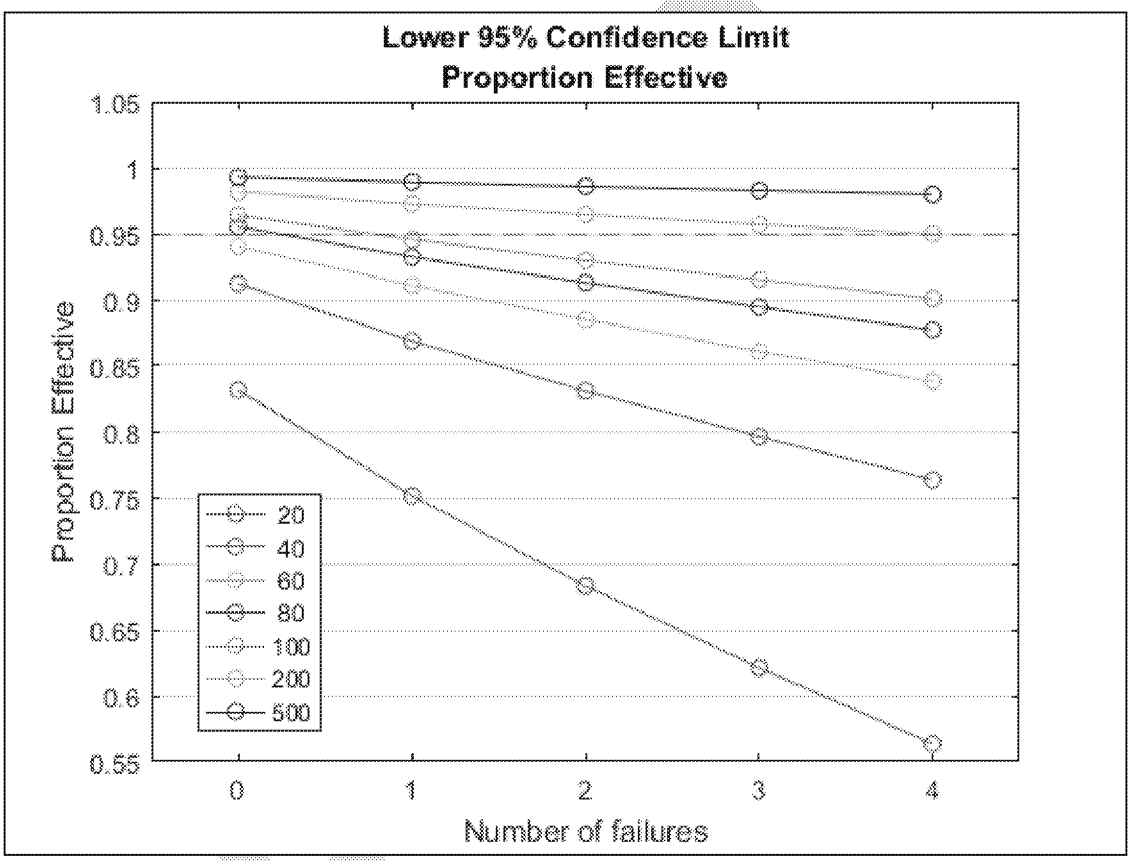

Figure [SEQ Figure * ARABIC]. Lower confidence limit for proportion effective vs number of failures and for a range of sample sizes.

1.3 References

McCullagh, P. and J. A. Nelder, 1983. *Generalized Linear Models*, Second Edition. Monographs on Statistics and Applied Probability: 37. Chapman and Hall,/CRC. NY.